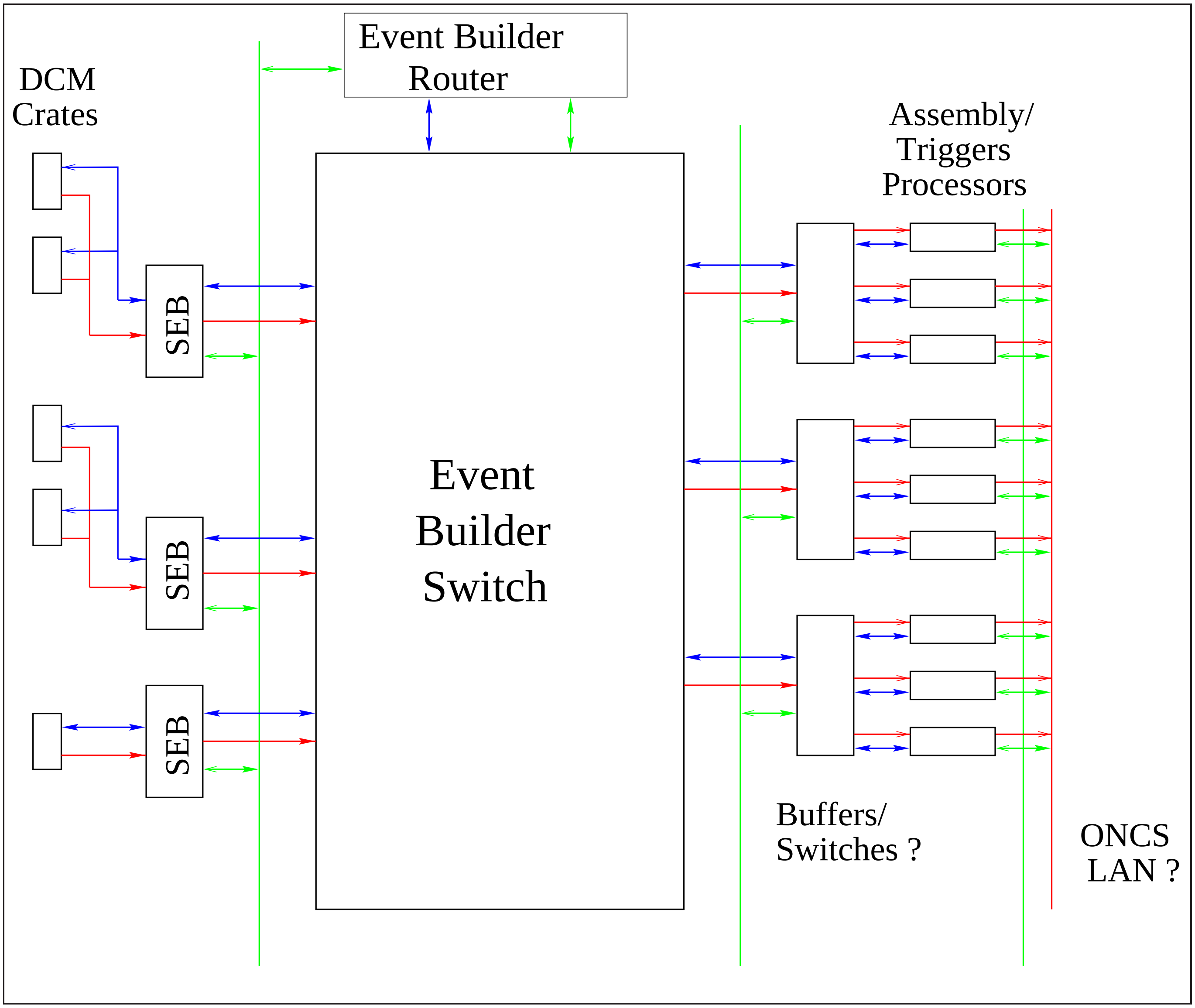
# The PHENIX Event Builder – Overview

## Major Functions

- Receive data from DCM's into sub-event buffers (SEB's).
- Switch fragments for given partition/crossing to single destination.
- Assemble events, run trigger algorithms in Assembly Trigger Processors (ATP's).
- Pass complete, trigger-selected events to ONCS.



# The PHENIX Event Builder – Overview (Cont.)

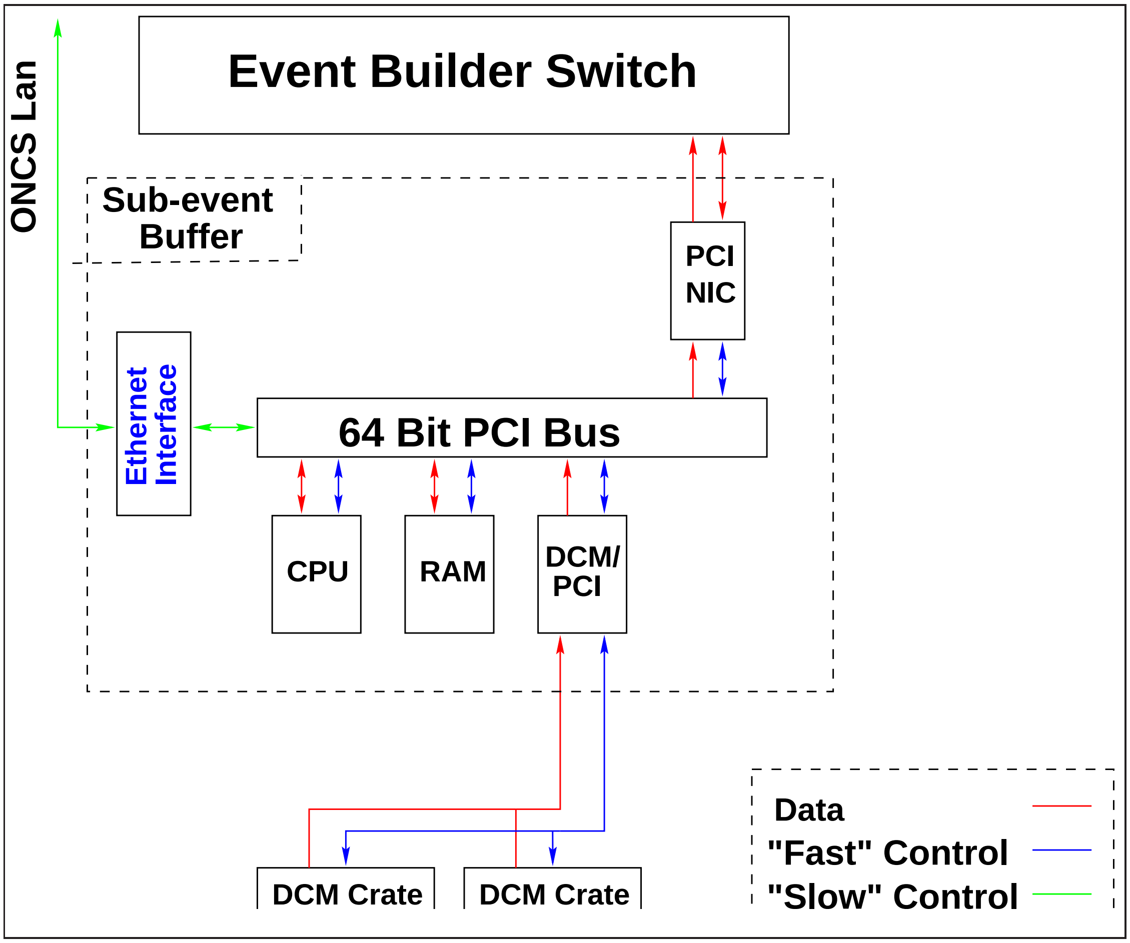
## Event-builder Components

- Sub-event buffers
  - Receives data from DCM's.
  - Determines destination(s) for received packet(s).
  - Provides framing for input to switch.
  - Sends data to switch.
- Router/Controller
  - Communicates with Assembly/Trigger Processor (ATP).
  - Provides fragment synchronization checking (?)
  - Receives destination requests from sub-event buffers.
  - Returns destination ATP to sub-event buffers.
  - Provides control/monitoring of sub-event buffers/switch.
- Switch
  - Receives data on $N$ inputs (input → sub-event buffer).
  - Routes data to $M(=N)$ outputs (output → Assembly/trigger processor.)
- (*name ?*) Assembly/Trigger processor
  - Receives data from switch.
  - Removes redundant/unnecessary framing.
  - Performs validity checks
  - Builds Pointer banks.
  - Runs level-(?) trigger algorithms.
  - Transfers event to ONCS.

# The PHENIX Event Builder – Sub-event Buffer (SEB)

## Major SEB Functions:

- Receive event fragments from DCM's.
- Buffer input to switch.
- Perform data integrity check (at some unspecified level).
- Determine/obtain event fragment destination.
- Send event fragments to switch.



# The PHENIX Event Builder – Switch input rates

## Available/standard transfer rates (< Gigabit)

- OC-1 - 52 Mbit/s
- 100BaseT Ethernet - 100 Mbit/s
- ATM UNI standard 100 Mbit/s
- OC-3 - 155 Mbit/s
- OC-12 - 622 Mbit/s
- Others (?)

## At 500 Mbyte/s event-building rate (neglecting overhead)

- 77 inputs at OC-1.
- 40 inputs at 100 Mbit.
- 26 inputs at OC-3.
- 7 inputs at OC-12.

## At 2 Gbyte/s event-building rate (neglecting overhead)

- 307 inputs at OC-1.
- 160 inputs at 100 Mbit.
- 103 inputs at OC-3.
- 26 inputs at OC-12.

## Comments

- DCM outputs at 50 Mbyte/s at full bandwidth →OC-12.
- Switch size becomes unwieldy for < OC-3.
- Ultimately, probably want to use OC-12 or greater.
- Cost of OC-12 currently probably too high.
- OC-3 or 100 Mbit initially OK.
- How to provide easy upgrade path ?

# Event Builder - Switch technology choice

## Considerations

- Intrinsic topology (e.g. LAN, Switch, HUB, Ring ...)
- Switch development.
- Standard link speeds.
- Link cost.
- Interface cost.
- Ease of use.
- Industry usage.
- Technology maturity.

## Possible Technologies

- ATM
  + Intrinsically point→point, Switch topology.
  + Gigabit switching speeds currently attainable.
  + Very high link speeds.
  + Heavy industry emphasis (Broadband ISDN)
  - Link cost - currently high.
  ? Ease of use ?
  - Immature technology - standards still developing.
- Fiber-channel
  + Intrinsically point→point, Switch topology.
  + Very high link speeds.
  + Mature technology.
  - Little industry implementation.
  ? Link cost ?
  ? Ease of use ?
  ? Switch development ?

# Event Builder - Switch technology choice (cont.)

## Possible Technologies (cont.)

- 100-BaseT (excluding 100-BaseT-VGany)
  - - LAN Topology, point-to-point usage possible.
  - - Slow link speeds (100 MBit/s max.)
  - ? Switch speeds ?
  - + Possibly heavy industry usage.
  - + Low Link cost.
  - ? Should be easy to use but for our application ?
- Custom Switch using Columbia QCD Nodes
  - + Can make specific to desired topology.
  - - Low link speeds.
  - - Custom technology.
  - + Low Link cost.
  - - Ease of use - requires top to bottom development.
  - -? "industry" usage at Columbia
- Cross-bar
  - + Most natural switch technology for event-builder.
  - ? Link speeds ?
  - + Switch speed essentially irrelevant.
  - - Industry cross-bar switches available, custom boards required.
  - + Low link cost.
  - - Ease of use - requires substantial development.
- Frame-relay (I am ignorant)

# Event Builder - Switch technology decision

## Current Status

- Two opens under serious consideration:
  - Top candidate - ATM.
  - Potential alternative - fast ethernet.
- How to proceed ?
- Vigorously pursue ATM
- Perform some tests of fast ethernet.
- Research the market – changing rapidly.
- Make event-builder design modular.

## Sources of expertise

- RD-31, built working prototype ATM event-builder.
- MIT CDF group (Paris Sphicas *et al.*), currently running 2x2 switch.
- CEBAF experiment (?) using ATM event-builder.
- Industry (BayNetworks, Fore, IBM, HP ...).

## Schedule

- I will be visiting RD-31, MIT group, BayNetworks, ... over next 2 months.
- RD-31 (Saclay group) has explicit proposal for participating.
- Have Nevis (and ONCS ?) participants attend formal training (?)
- Major decisions/milestones
  - Technology choice (ATM vs Ethernet) - June 1.
  - Hardware vendor(s) - June 1.
  - Connect (4 ?) processors through switch - Sep 1.
  - Make minimal SEB+switch+ATP system functional - Jan 1, 1998.

# The PHENIX Event Builder – Buffering

## Buffering requirements

- Actual requirements at sub-event buffer and ATP's unknown.
- Requires study of switch performance and trigger algorithms.
- Assume central Au-Au has twice current (old) average event size - 400 kbyte.
- Educated guess - Sub-event buffers
  - Worst-case (?), on-average SEB sees 1/10 of full event size (40 kbyte)
  - Suppose we want 20-event deep buffer.
  - Need 2 Mbyte buffer – clearly not a problem.
  - Necessary but not sufficient – need to be able to hold non-zero supressed event.
- ATP buffering
  - Suppose 4 Mbyte available per node.
  - 10 event-deep buffering per node.
  - Likely to have 50 nodes → 500 event buffering capacity.
  - Even with poor utilization this should be sufficient.

# The PHENIX Event Builder – Buffering

## Where to buffer ?

- (DCM output ports)
- Sub-event buffers
  - Absorbs fluctuations in front of switch.
  - Allows control of data rate into switch.
  - Back-pressure exerted on DCM outputs.
- Assembly/Trigger processors.
  - Use memory in ATP's to absorb processing rate fluctuations.
  - Distributed buffering system – no "clogging" by full buffers.
  - Back-pressure (re-direction) exerted through router.
- **OR Do we need buffers between Switch/ATP's ?**

  **- In principle not necessary – switches sufficient.**<br>
  **- Add significant cost (or custom hardware).**<br>
  **+ Buffers allow ATP's to be decoupled from switching.**<br>
  **+ Buffers reduce memory needed in ATP's.**<br>
  **+ Buffers provide more local routing of events.**

# The PHENIX Event Builder – Routing Schemes

## Level-1 determined routing (deprecated)

- Determine event destination at Level-1.
- For a given run allocate ATP nodes per partition.
- Use pre-determined scheduling algorithm.

+ Simple, deterministic algorithm.

- Non-adaptable to congestion, ATP failure.
- Very un-modular.

## Deterministic Switch router

- Route determined at sub-event buffer .
- Deterministic algorithm using pre-allocated nodes per partition.

+ Simple algorithm – posibly first implemented.
+ Very modular, only router and switch knows about route addressing.

- Non-adaptable to congestion, failure.
- Static allocation of nodes may not be optimal.

## Adaptable Switch router

- Route determined at sub-event buffer .
- Feedback from ATP's used to make routing decision.
- Many possible ways to implement.

+ Very modular, only router and switch knows about route addressing.
+ Adaptable to congestion, failure.
+ Provides dynamic re-allocation of nodes.

- More complicated algorithm (feedback problems ?)
- Requires communication with ATPs.

# The PHENIX Event Builder – Router Implementation

## Issues

- SEB ↔ Router connection must have small ($< 5 - 10\mu s$) latency.
- Router must be able to address all SEB's simultaneously (broadcast ?)
- Communication with router should be robust use simple protocal.
- Communication with router should be unaffected by data rate (?)

## How to connect router to SEB's ?

- Have router, SEB's reside in VME.
- Connect SEB's, router with cable bus/LAN.
- Connect SEB's, router using ATM but external to data switch.
- Connect SEB's, router through data switch.

## Possible routing/data integrity verification algorithms

- Minimal interference
  - First fragment for (partition, event) iniates routing decision.
  - Decision does not incorporate event size.
  - Result broadcasted to all SEB's.
  - All fragments report to router.
  - No integrity checking.
- Maximal interference
  - Router waits for all fragments.
  - Router makes decision (using event size ?).
  - Result broadcasted to relevant SEB's.
  - Events with missing fragments marked/dropped.
- Intermediate solution
  - All fragments report to router.
  - Router makes decision on first fragment.
  - Broadcasts result to SEB's.
  - Forwards fragment list/expected event size to ATP.

# Event Builder - Level-2.5/ONCS Interface

## Previous Discussions/Decisions

- Level-2.5/OCS Interface resides in Level-2.5 processors.
- Accepted event passed to ONCS interface code.
- Event allowed to be passed in non-contiguous fragments.
- No significant re-formatting done in Level-2.5 processor.

## Implications/Considerations

- Desire zero (minimal)-copy transfer from input-output.
- Data altered mainly through framing removal and pointer bank, trigger primitive addition.
- Assembly/trigger processing/ONCS interface code share memory.
- Need appropriate memory management algorithm.
- How to handle multiple tasks in same processor ?

## Control/Operation

- How to provide feedback to router ?
- Can one processor handle multiple partitions ?
- How to decide when received event is complete ?
- Must event ordering be maintained ?
- Time-out mechanism needed for trigger calculation ?
- How to prevent memory lock-up: trigger requires more space for output than available, processor stuck.

# Sub-event Buffer (SEB) – Attack Plan

## Requirements

- Proto-type of SEB available for Fall Sector test.
- Same design useful for Phenix running without major mods.
- SEB design satisfy requirements for first (2 ?) years Phenix operation.
- SEB design must accomodate Switch technology decision.

## Division of reponisibilities

- BNL - Data input
  - Finalize DCM output protocal - **Nevis/DCM, John**.
  - Design/construct DCM → PCI interface - **John**.
  - Provide DCM → PCI driver - **John**.
  - Provide Buffer reading/management software - **ONCS**.
  - Provide control interface to DCM → PCI interface - **ONCS**.
  - Provide control interface to buffer manager - **ONCS**.
- Nevis - Data output
  - Provide switch network interfcace card (NIC) - **Nevis/Evb**.
  - Provide Switch NIC drivers - **Nevis/Evb**.
  - Provide fragment routing/control software - **Nevis/Evb**.
  - Provide control interface to fragment routing/control - **Nevis/Evb**.
  - Provide control interface to switch NIC - **Nevis/Evb**.
- Joint - Hardware/Operating system decision.
- ONCS/Event-builder interface takes place in buffer manager.

## Schedule

| Task | Target Date |
|---|---|
| Finalize DCM output protocol | Jan. 31 (?) |
| Choose initial hardware | Feb. 11 |
| Choose operating system | Feb. 11 |
| Produce prototype DCM →PCI interface | Summer |
| Write DCM →PCI interface device driver | Summer |
| Decide switch network technology | Jun 1 (?) |
| Obtain switch network interface | Jun 1 (?) |
| Implement switch NIC driver | Jul 1 (?) |
| Implement CORBA interfaces | ? |
| Test Seb throughput | Sep 1 (?) |

# Event Builder - Manpower

## Task Summary

- Output half of SEB's.
- Switch implementation, control, ...
- Router implementation.
- Input half of ATP's.
- Control interfaces.
- Test, monitoring system.

## Manpower situation at Nevis

- Currently have available/expect
  - Myself
  - Bill
  - Jamie Nagle
  - Another post-doc
  - 2 graduate students over summer
- Clearly not enough !
- What do I think we need (in addition to above) ?
  - Additional engineering support.
  - Additional post-doc ?

## Other sources of manpwer/support

- RD-31
- BNL (ONCS)
- Industry (consulting ?)
- Columbia/other computer science (?).